Accepted Manuscript

Saer Samanipour, Jake W. O'Brien, Malcolm J. Reid and Kevin V. Thomas. 2019.
Self Adjusting Algorithm for the Nontargeted Feature Detection of High Resolution Mass Spectrometry Coupled with Liquid Chromatography Profile Data.
Analytical Chemistry . 91 (16): 10800-10807.

A Self Adjusting Algorithm for the Non-targeted Feature detection of High resolution mass spectrometry coupled with liquid chromatography Profile Data

Saer Samanipour,*,†,‡ Jake O´Brien,‡ Malcolm J. Ried,† and Kevin V. Thomas†,‡

†*Norwegian Institute for Water Research (NIVA), Gaustadalléen 21, Oslo 0349, Norway*

‡*Queensland Alliance for Environmental Health Sciences (QAEHS), The University of Queensland, 20 Cornwall St, Woolloongabba, Qld 4102, Australia*

E-mail: saer.samanipour@niva.no

Phone: +47 98 222 087

Abstract

Non-targeted feature detection in data from high resolution mass spectrometry is a challenging task, due to the complex and noisy nature of datasets. Numerous feature detection and pre-processing strategies have been developed in an attempt to tackle this challenge, but recent evidence has indicated limitations in the currently used methods. Recent studies have indicated the limitations of the currently used methods for feature detection of LC-HRMS data. To overcome these limitations, we propose a self adjusting feature detection (SAFD) algorithm for the processing of profile data from LC-HRMS. SAFD fits a three dimensional Gaussian into the profile data of a feature, without data pre-processing (i.e. centroiding and/or binning). We tested SAFD on

55 LC-HRMS chromatograms from which 44 were composite wastewater influent samples. Additionally, 51 of 55 samples were spiked with 19 labeled internal standards. We further validated SAFD by comparing its results with those produced via XCMS implemented through MZmine. In terms of ISs and the unknown features, SAFD produced lower rates of false detection (i.e. $\leq$ 5% and $\leq$ 10%, respectively) when compared to XCMS ($\leq$ 11% and $\leq$ 28%, respectively). We also observed higher reproduciblity in the feature area generated by SAFD algorithm versus XCMS.

Introduction

High resolution mass spectrometry coupled with liquid chromatography (LC-HRMS) is one of the main analytical tools for analysis of small polar and semi-polar organic compounds in complex samples, with application in the areas of pharmaceutical development, human health, metabolics and environmental monitoring (to name just a few).[1–8] Chemical identification is commonly performed through a combination of target, suspect, and non-target analysis.[5–8] Target and suspect screening approaches focus on a limited number of well-known chemicals and they are considered relatively reliable and accurate in the identification of organic compounds in complex samples.[1,8–11] On the other hand, non-target analysis (NTA) aims at simultaneous identification of known and unknown organic chemicals in the samples, using the data generated by LC-HRMS[1–4,12,13] without prior knowledge regarding the non-target analytes.

Feature/peak detection is one of the most crucial steps in non-targeted LC-HRMS workflows from both qualitative and quantitative points of view.[14–16] During feature detection, data complexity is reduced from $\approx 1{\times}10^{+8}$ variables to $\leq$ 10,000 features/peaks through grouping of the related signals (i.e. all masses measured within a feature/chromatographic peak).[2,3,13] The generated lists of features are then used as inputs to chemical identification workflows.[1–3,13] However, the noisy and complex nature of HRMS data means that current

feature detection strategies are prone to error, and these errors result in lower levels of reproducibility and robustness.[4,17]

There are different open access/source algorithms for the feature detection of LC-HRMS data such as XCMS[18] and MZmine.[19] Although there are numerous differences between the algorithms, they do share a common framework around the use of 2 dimensional data (i.e. centroided data[2,3]) rather than 3 dimensional data (i.e. profile data) and the use of extracted ion chromatograms (e.g. XICs and/or region of interest[3,16]). These approximations are made in order to reduce data size and consequently decrease the data processing time, but they come at the cost of the necessity for a suite of optimizable parameters that the users need to carefully set in order to minimize the rate of false detection.[20,21] However, multiple studies have shown that the feature detection using this procedure, even under optimized conditions, is prone to high rates of false detection.[22–25] As of today, there have been only a few studies working with the three dimensional (3D) data.[26,27] One such method used a probabilistic approach,[27] while the other one employs the artificial neural networks for the feature detection in the LC-HRMS data.[26] The main disadvantages of these methods are the fact that they need to be trained and in the case of artificial neural networks the data needed to be binned prior to their use.

In this study, we present a self adjusting feature detection algorithm (SAFD) that utilizes all of the points measured within a feature without data centrioding and data binning. This algorithm is considered self-adjusting due to the fact that it utilizes user defined parameters as only the first guess in an adoptive process. SAFD fits a 3D Gaussian distribution into the profile data generated via LC-HRMS to detect features. The proposed algorithm does not need optimization of parameters such as the peak widths in mass and time domain in the same way as previously reported methods. SAFD was tested and validated using a dataset of 55 LC-HRMS chromatograms including 44 wastewater influent samples spiked with 19

internal standards (IS). Furthermore, we validated SAFD by comparing its feature lists with those generated via XCMS implemented through MZmine.

Experimental Section

The Experimental Setup

In total 55 samples consisting of 4 blank, 4 equilibration injections, 3 internal standard injections, and 44 composite wastewater influent samples (Section S2) were analyzed using LC-HRMS. All the samples except the 4 equilibration samples were spiked with 19 labeled internal standards (IS) at 10 ngL^{-1} of each standard, Table S1. In this study we looked at the rates of false detection both among ISs and overall detected features. In the case of ISs, the spiked samples were used for evaluation of the true positive and false negative detection while the 4 equilibration samples were used for false positive detection evaluation. We refer to a feature that its presence confirmed (i.e. a true peak) in a sample as a true positive (TP) and a feature that its absence is confirmed in the sample as a true negative (TN). A false negative (FN) is a case where a TP is not detected by the tested method whereas a false positive (FP) is a TN identified as a feature by the algorithm.

Sample Preparation and Analysis

All the samples were filtered and transferred into 1.5 mL vials with a total volume of 1 mL (more details are available in Section S2 of the Supporting Information). All the samples, including the blanks, were then spiked with the mixture of ISs and were stored in freezer until the analysis. A detailed list of solvents, ISs, and their supplier is provided in the SI, section S1.

All the samples were analyzed on an AB Sciex 5600+ QToF (Sciex, Concord, Ontario, Canada) LC-HRMS. We, directly, injected 10 μL of each sample into the instrument without

any other sample preparation step. For more details regarding the instrumental conditions, please see Section S3.

IS Identification

For IS detection and identification, we employed a semi-targeted approach where we first performed a non-targeted feature detection and then the feature lists were searched for the ISs. For ISs to have their presence confirmed in the samples, they had to have a mass error $\leq$0.003 Da and a retention error $\leq$10 seconds. This approach was previously shown to be effective for identification of target analytes in complex environmental samples.[28–30]

Self Adjusting Feature Detection Algorithm (SAFD)

All the raw chromatograms were converted into an open ms format (i.e. mzXML)[31] via MSConvert provided by the ProteoWizard package.[32] The converted chromatograms were processed employing the self adjusting feature detection algorithm (SAFD) in order to detect all chromatographic features in the data, which had an intensity larger than the user set threshold (Table S2). This algorithm is an iterative one where the features are processed one at the time starting with the feature with the highest intensity. Once a feature is detected in a chromatogram, the signal of that feature is set to zero and SAFD moves forward with the detection of the next most intense feature in the sample. The SAFD goes through 9 steps during each iteration (i.e. detection of a feature in the chromatogram). These steps are: 1) maximum detection, 2) half-height placement (mass domain), 3) signal smoothing, 4) signal interpolation, 5) Gaussian fit (mass domain), 6) baseline tracing, 7) move to the neighboring scans, 8) Gaussian fit in time domain, and 9) removal of the signal of the detected feature.

Maximum detection and half-height placement (steps 1 and 2): After finding the most intense location in the chromatogram (Fig. S1), the half-height of that mass peak is calculated by dividing the intensity of the apex by two. In order to locate the peak half-

height in the data, a mass window is calculated employing the user defined mass resolution (i.e. first guess) and the mass of the apex. In the next step the intercepts between a line spanning within the calculated mass window at the level of the apex half-height and the measured signal are found (Fig. S2). The found intercepts enable us to define the true mass window (i.e. peak width in the mass domain) and the resolution based on the experimental data. This signal (i.e. above apex half-height) and the measured parameters (i.e. true mass window and the resolution) are used in the next steps of the feature detection.

Signal smoothing (step 3): The recorded signal in the previous step (i.e. above apex half-height), then goes through a smoothing step. This step reduces the levels of signal fluctuation before performing the signal interpolation. For the smoothing step a simple moving average with an averaging window of three points are used (Fig. S3). The milder smoothing method (compared to Savitzky Golay methodology) and a small averaging window were necessary for minimizing the signal alteration while reducing the signal fluctuations.

Signal interpolation (step 4): The smoothed signal is interpolated using the spline function[33] with a total number of 50 points. This step generates two vectors of 50 points each for masses and intensities, respectively (Fig. S4). The signal interpolation is a necessary step in SAFD due to the fact that, depending on the instrumental resolution, there are not enough measured points in the top 50% of a mass peak for fitting a three parameter Gaussian.

Gaussian fit (step 5): The interpolated data is used for fitting a three parameter Gaussian function, (Fig. S5) where A is the signal amplitude (i.e. the signal intensity at apex), σ is the measured mass window during step 2 (i.e. half-height placement), and μ is the measured mass of the apex.

$$f(x, A, \mu, \sigma) = \frac{A}{\sigma\sqrt{2\pi}} \exp \frac{-(x-\mu)^2}{2\sigma^2}. \tag{1}$$

Once the interpolated signal is fitted using the Gaussian function via a least square method,[34] the algorithm produces a regression coefficient (i.e. R^2) for the goodness of the fit and the model estimation of the three parameters of the Gaussian function, Eq. 1. The regression coefficient is employed as a means to acceptance or rejection of the fit, by comparing it to a user defined threshold (the default of 0.9). SAFD utilizes the top 50% of a mass peak for Gaussian fitting, in order to minimize the influence of the neighboring mass peaks, which increases the accuracy of this algorithm.

Baseline tracing (step 6): At this stage, the Gaussian model is extrapolated to reach the baseline (i.e. the user defined minimum intensity). Doing so enables the definition of the mass window in which the baseline must be found. To find the baseline a similar approach to the half-height placement (i.e. step 2) is used, where the intercepts of a line at the level of baseline lying within the defined mass window and the measured signal are measured (Fig. S6). Once the measured baselines are found, all the masses and intensities within the boundaries of the mass peak baselines are recorded for the peak integration. At this point the algorithm has collected all the necessary information regarding the detected mass peak.

Neighboring scans (step 7): After the detection of the center mass peak the SAFD moves in the time domain by repeating the process between step 2 (i.e. half-height placement) and step 6 (i.e. baseline tracing) for the neighboring scans. During this process the algorithm uses the measured resolution for the previous mass peak (i.e. scan number -1) rather than the user defined one (i.e. first guess) for defining the mass peak boundaries. The algorithm moves away from the center mass peak in both directions (i.e. the scans larger and smaller than the center peak) until it receives the stopping signal (Fig. S7). The stopping signals consist of three different user defined threshold, which in case of violation the algorithm stops the mass peak detection process (i.e. moving in the time domain). These thresholds include R^2, minimum intensity, and minimum signal increment. In case of R^2, if the calculated regression coefficient for a mass peak is smaller than the user defined threshold the algorithm assumes

that the signal in that scan is not a real signal but noise. Therefore, it stops the mass peak detection. Another stopping signal is issued if the apex intensity of the next scan is smaller than the user defined minimum signal intensity. Finally, the SAFD algorithm assumes that within a chromatographic peak (i.e. in time domain) as you move away from the apex the signal intensity should be smaller than the previous scan. Consequently, an increase in the signal may indicate the presence of overlapping peaks. Therefore, if the algorithm observes such a trend, it stops the mass peak detection assuming the presence of an overlapping peak in the time domain.

Gaussian fit in the time domain (step 8): Once the algorithm receives the stopping signals in both directions (i.e. the scans larger and smaller than the center peak), it fits a three parameter Gaussian function into the recorded signal in the time domain (Fig. 1). If the Gaussian fitting process is successful (i.e. R^2 larger than the set threshold), the algorithm considers that as a successfully detected feature and calculates the average mass, retention time, minimum measured mass, maximum measured mass, minimum retention time, maximum retention time, feature height, feature area, and the average feature resolution, based on all the recorded points within that feature. All the mentioned recorded information is reported in the final feature list. It should be noted that the overall process carried out during the feature detection is equivalent of fitting a 3 dimensional Gaussian[35] into the measured signal.

Signal removal (step 9): Once all the information regarding a chromatographic feature is recorded, independently from its successful detection, its signal is set to half of the user defined minimum intensity (Fig. S8). This step enables the algorithm to detect the next most intense feature in the sample without the interference of the already processed features.

It should be noted that the SAFD algorithm does not distinguish between the features related to a chemical component and potential adducts, isotopes, and/or in-source fragments.

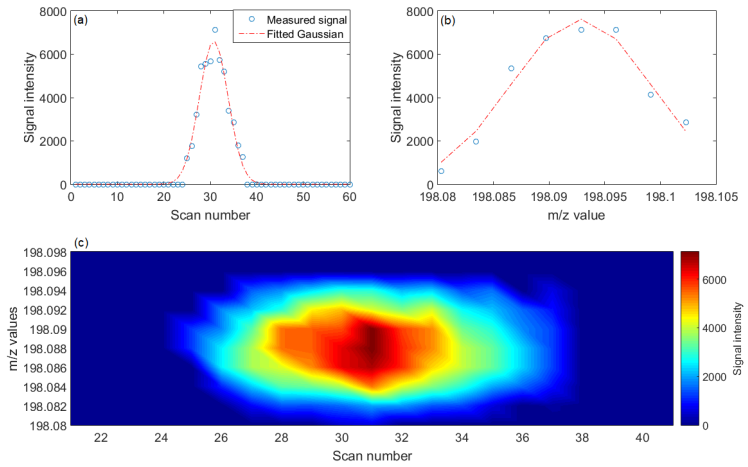


Figure 1: Depicts (a) the fitted Gaussian in the time domain, (b) the fitted Gaussian on the base peak in the mass domain, and (c) a contour plot of the detected feature, step 8. The presented plot is based on a feature of caffeine (IS) in the wastewater sample.

Consequently, each of these signals will be detected as an individual feature. Therefore, the analyst, if deemed necessary, must filter the feature lists for the removal of the potential adducts, isotopes, and/or in-source fragments.

SAFD Parameters

The algorithm takes four types of inputs: importing parameters, stopping parameters, filtering parameters, and performance essential parameters, Table S2. The importing parameters include path to the file, the file format, and finally mass range limit (if necessary). As for stopping parameters, they consist of four thresholds to stop the algorithm from moving forward in the time domain. These thresholds are related to: R^2 (i.e. mass domain regression coefficient set to 0.9), maximum signal increment to avoid grouping the overlapping features (defined at 5%), minimum intensity of the mass peak (set to 2000 counts), and maximum number of iterations (defined at 15,000). The filtering parameters, i.e. minimum peak width (2 seconds) and maximum peak width (300 seconds) in time domain, are used to remove the time domain features that are considered noise/background from the feature lists (i.e. very broad peaks). Finally, the performance essential inputs are the mass resolution and

the minimum peak width in the mass domain. These two parameters are not completely independent from each other. The user defined mass resolution parameter is utilized as the initial value for defining the peak width in the mass domain. For the masses smaller than 200 Da, the LC-HRMS instruments have lower resolution compared to the larger masses. The parameter minimum peak width in the mass domain is set to deal with this issue. In other words, if the defined mass window based on the user set resolution is smaller than the minimum peak width, the algorithm adjusts the resolution in order to produce a peak width equal to minimum peak width.

Two stopping parameters related to R^2 and maximum signal increment are purely connected to the nature of the signal and are dataset independent. Therefore there is no need for their optimization. For the minimum intensity of the mass peak and maximum number of iterations, these parameters depend on the complexity of the analyzed samples. Therefore, the analysts must use prior knowledge to define these parameters. For example, our previous experience with wastewater samples and LC-HRMS[10,13,29] indicated that a maximum number of detectable features (i.e. iterations) and minimum signal intensity of 15,000 and 2,000, respectively are adequate for these types of samples. The same approach was used for the two filtering parameters of minimum peak width and maximum peak width in time domain.

Finally for performance related inputs of the mass resolution and the minimum peak width in the mass domain, we optimized them, by evaluating them for randomly selected 10 ISs in 5 different samples. We employed the average observed resolution (in this case 20,000 half width full-scan) and minimum peak width of 0.02 Da as the optimized settings for these parameters.

XCMS via MZmine Parameters

In order to validate the SAFD algorithm, we compared its feature list with the one produced by XCMS feature detection algorithm[18] implemented via MZmine[19] and RCall package. XCMS was selected due to its wide use and the fact that it is extensively documented.[22–24] For the common parameters between XCMS and SAFD algorithms, we used the same settings whereas for the parameters specific to XCMS, we employed the average values defined based on the features used for SAFD optimization and the preview function implemented in MZmine. The list of all the parameters and their settings is provided in Table S3.

Calculations

All the calculations were run using a work station with 12 CPUs and 128 GB of memory. SAFD algorithm is developed employing julia 1.03 programing language.[36] All the figures are generated using the matplotlib[37] (i.e. developed within python 3[38]) and PyCall modules. All the functions and scripts will be made available as a julia package with MIT license through GitHub. Prior to the package release, the scripts/functions are available under MIT license upon request.

Results and Discussion

All 55 chromatograms were processed with both SAFD and XCMS (via MZmine). We compared the unique feature lists via these algorithms to each other with a particular focus on the ISs. The performance of the methods was compared by evaluating feature detection through the rate of false detects as well as the reproducibility of integration. Finally, the sensitivity of SAFD as a function for the two performance affecting parameters (i.e. the mass resolution and the minimum peak width in the mass domain) was assessed.

Feature Detection

All 55 chromatograms were processed using the SAFD algorithm and employing the optimized parameters. SAFD produced a feature list for each chromatogram reporting the average mass, scan number, retention time, minimum measured mass, maximum measured mass, minimum retention time, maximum retention time, feature height, feature area, and the average feature resolution. These feature lists were then combined to generate a master feature list via SAFD taking advantage of a home-developed alignment function[39], that uses the individual feature information for the alignment. The MZmine master feature list was generated, using the feature alignment function implemented in MZmine with a mass window of 0.01 Da and retention window of 0.2 minutes. The absence and/or presence of each IS was manually checked in the samples and compared to the master feature lists generated by the tested algorithms to assure that the false detection cases are not caused by mis-alignment. Based on the results of the ISs, both alignment algorithms were successful in generation of the master feature lists.

SAFD produced 3445 unique features in all 55 chromatograms whereas XCMS via MZmine detected 3273 unique features in the same samples. Among the detected features, both methods detected 2032 (59%) whereas 1413 features were only detected by SAFD and 1241 were detected only by XCMS via MZmine. To evaluate the overall rate of false positive detection for each method, we randomly selected 50 detected features in three samples (i.e. $3\times50=150$) from each of the three groups (i.e. only SAFD, only MZmine, and both) for further evaluation. For the selection criteria of FP features, we employed the method suggested by Myers et al.,[25] which consisted of manual inspection of the features to the expected feature shape (i.e. a Gaussian). Among the features detected by both methods, we found only three cases of FP detection. On the other hand, for the method specific features, SAFD algorithm produced 14 FPs while XCMS via MZmine resulted in 42 cases of FP detection. In addition to those evaluated cases, we further examined all the IS features (i.e. total of $55\times19=1045$

detection cases) in the samples for false detection rate. The well-known nature of those features enabled us to evaluate the reason behind the observed false detection cases.

For the ISs, SAFD algorithm resulted in 26 cases (i.e. $\leq 5\%$) of false negatives (FNs) whereas the XCMS via MZmine produced 117 cases (i.e. $\leq 11\%$) false negative detections, Fig. 2. None of the methods produced any cases of false positive for the ISs. Among the 26 FN cases of SAFD algorithm, 22 were caused by the minimum intensity threshold of 2000 counts, Fig. S9. The remaining four FNs, were caused by the stopping parameter maximum signal increment. For these four cases, the signal had a high level of noise in the time domain, which stopped SAFD prematurely, Fig. S10. Consequently, those features did not meet the filtering parameter of minimum peak width of 2 seconds and therefore they were not detected. Our investigation in the FN cases that were specific to XCMS via MZmine (i.e. 117-22=95 cases) appeared that all 95 FNs were cases where the peak is present in the XIC of the ROIs however, during the feature detection the CentWavelet algorithm was not able to detect these features. This detection failure could be caused by a variety of reasons, including the five internal filters on the XICs before sending them for feature detection or the feature detection algorithm itself.[25] We modified the three parameters related to CentWavelet algorithm (i.e. signal/noise and Wavelet scales). The changes in the signal/noise and upper limit of Wavelet scales did not result in any improvement in the detection of the missed features (i.e. FNs). On the other hand the changes in the lower boundary of the parameter "Wavelet scales", from 0.1 to 0.2 minutes, caused the positive detection of hydroxycotinine in sample 5 while resulting in a FN for the same IS in sample 8. This suggests that further investigation of the effect of each parameter on the performance of the XCMS feature detection is needed.

SAFD algorithm is effective in the detection of features in the LC-HRMS data of wastewater influent samples. This algorithm appeared to perform better than the XCMS algorithm

in minimizing the false discovery rate (i.e. FP and FN detection cases). Additionally, the parameter setting of SAFD algorithm is very simple and intuitive.

Feature Integration

We also compared the performance of the SAFD algorithm and XCMS implemented via MZmine in feature integration. Both algorithms produced area and height for each detected feature. The quality of integration is highly crucial to both non-target analysis and omics experiments, especially if the feature prioritization is done through statistical approaches. In this case also, we focused on the features of ISs, given the total number of unique features in all the samples (i.e. $\simeq$ 3,000). Considering that all the samples, except the equilibration injections, were spiked with the same amount of ISs, we utilized the observed variability in the feature areas across the samples as an indication for the quality of integration.

The SAFD algorithm consistently resulted in lower averaged absolute standard error of integration for all the ISs in the spiked samples. The averaged absolute standard error of integration for SAFD algorithm was 20% whereas for XCMS method this error appeared to be of 57%, Fig. 3. We further compared these calculated standard errors using the non-parametric test Kruskal-Wallis test.[40] A ρ value ≤ 0.01 suggested the rejection of the null-hypothesis and that these two error sets are statistically different from each other. Moreover, examination of the variance of the standard errors, Fig. S11, and the standard deviation of the averaged standard errors further indicated the overall superior performance of SAFD algorithm in the feature integration compared to the XCMS algorithm. SAFD algorithm for two out of 19 ISs resulted in a significantly larger than average standard error, Figs. 3 and S11. These ISs, atrazine desisopropyl and atrazin desethyl, both were in the middle of the chromatogam (i.e. retention times $\simeq$ 5 min) and consistently generated larger feature areas in blanks compared to real samples. These suggest the presence of matrix effect manifested in higher ion suppression for real samples compared to blanks.

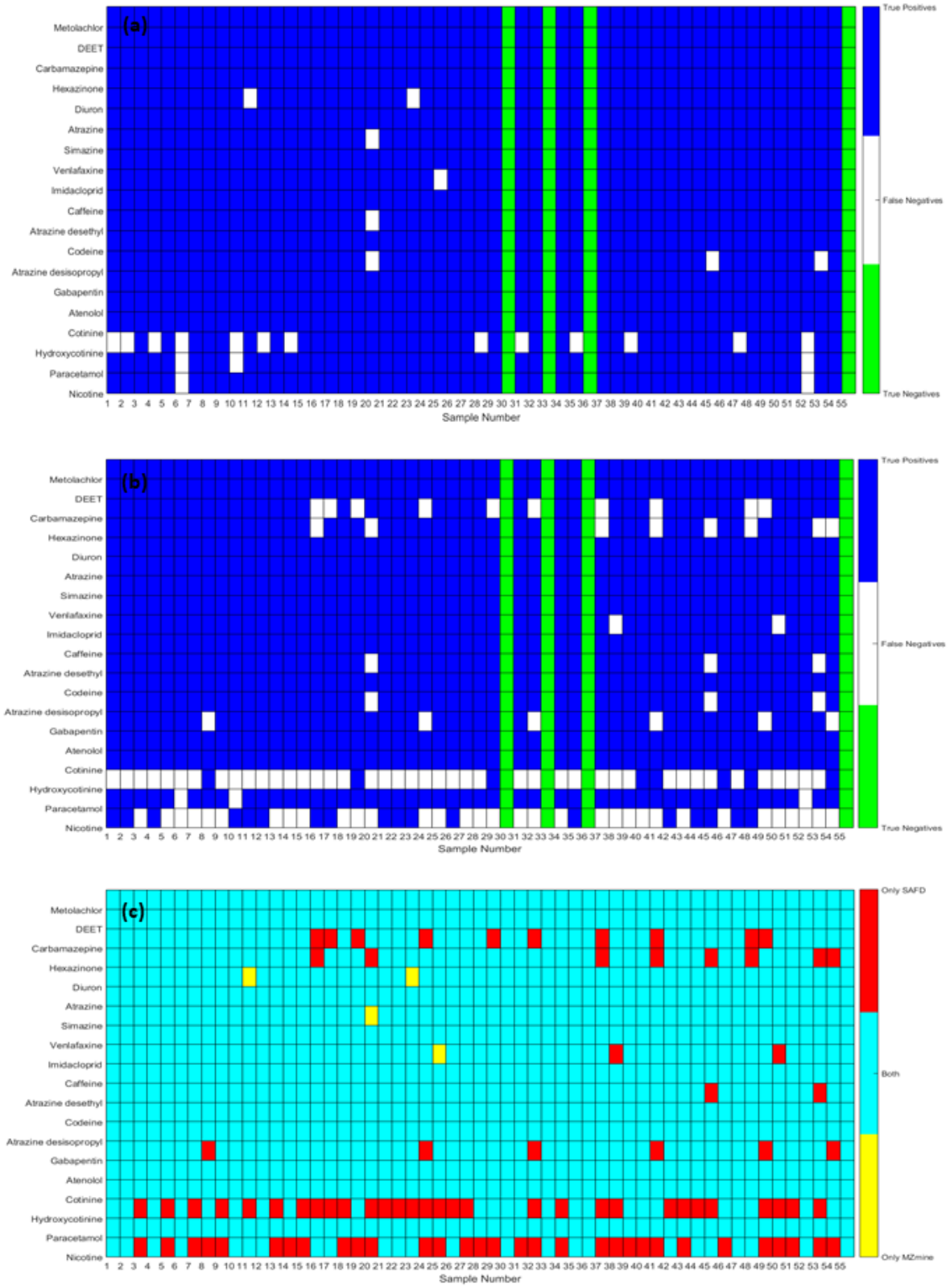


Figure 2: Depicting the detection matrix of ISs via (a) SAFD algorithm, (b) XCMS via MZmine, and (c) the difference between the two algorithms.

We observed a high level of linearity between the feature heights and area for the ISs via both algorithms with Pearson correlation[41] coefficients of $\simeq 0.85$, Fig S12. The high correlation coefficients indicate a direct correspondence between the feature heights and feature areas. The feature areas calculated by SAFD algorithm appeared to be one order of magnitude larger than those ones via XCMS. This discrepancy is related to the way that feature areas are calculated by each method. It should be noted that the trends/relative values for feature areas are far more significant than the absolute values.

The developed feature detection algorithm (i.e. SAFD) successfully integrated all the detected features across all the spiked samples keeping the standard averaged standard error within the acceptable experimental error (i.e. 20%). The cases where the observed standard errors were significantly larger than 20% appeared to be caused by the background effect through ion suppression. Overall, SAFD algorithm appeared to perform better than XCMS algorithm in accurately integrating the features in the analyzed samples.

Sensitivity Analysis

We evaluated the sensitivity of the SAFD algorithm towards the two performance essential parameters (i.e. mass resolution and the minimum peak width in the mass domain). It should be noted that, normally, these two parameters are not independent. In the SAFD algorithm, the minimum peak width is introduced to handle exceptions, where the first estimate of the mass resolution (i.e. the user defined value) results in too small of a mass window. To test the algorithm's sensitivity towards these parameters, we randomly selected 10 IS features in 5 samples (also randomly selected) and integrated those features setting the mass resolution ranging between 5,000-85,000 (six steps) and varying the minimum peak width from 0.001 Da to 0.08 Da (seven steps). The average integration error of the 10 features assuming SAFD results under the optimized conditions (i.e. the resolution of 20,000

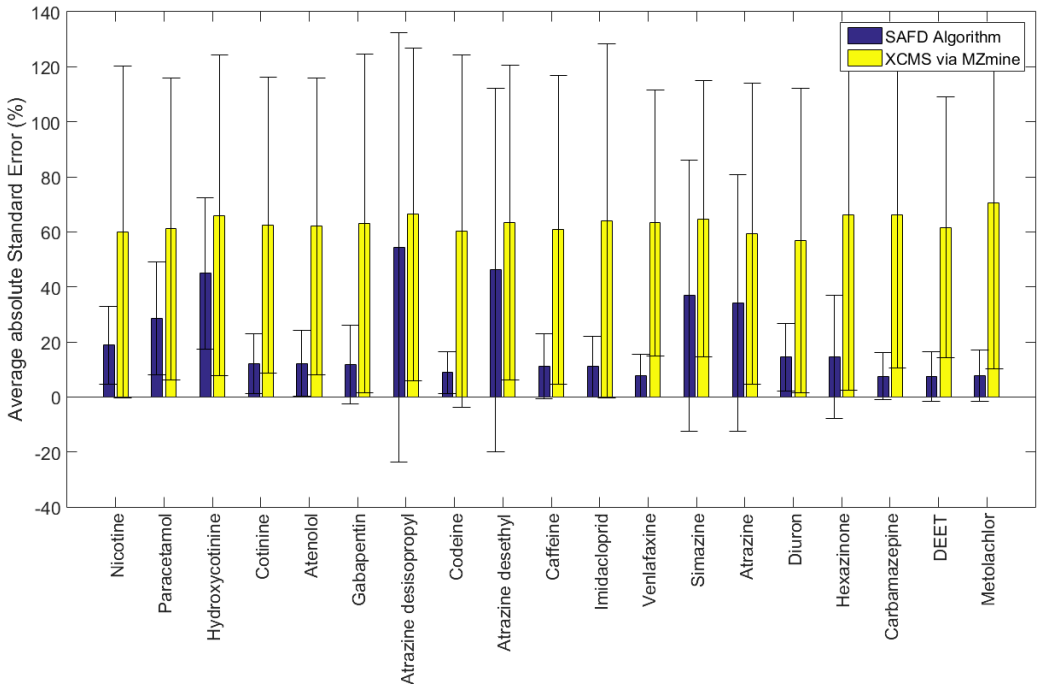


Figure 3: Shows the calculated average absolute standard error for each IS over 51 spiked samples. The error bars depict the calculated standard deviation.

and minimum peak width of 0.02 Da) as the truth was calculated for each point in the grid.

The algorithm appeared to be sensitive to extreme cases where both parameters are set wrongly (i.e. too far from the optimized conditions), particularly for the mass resolution, Fig. S13. The results of our sensitivity analysis indicated that for resolutions $\leq$ 10,000, SAFD algorithm is more prone to produce non-optimized results. For the mass resolutions $\geq$ 15,000, the minimum peak width may range between 0.010 and 0.05 Da without affecting the performance of SAFD algorithm, Fig. S13. In the case of extremely high resolution settings (i.e. $\geq$ 35,000 for this dataset) the algorithm systematically ignored the set resolution and treated that detection case as an exception. Consequently, the algorithm used the defined minimum peak width rather than the set mass resolution. It should be noted that only under the resolution setting of 85,000 and the peak width setting of 0.001 Da, SAFD produced five cases of FNs, which further indicates the robustness of the algorithm.

Overall, SAFD appeared to be very robust and highly stable during the sensitivity analysis. The lower sensitivity of the algorithm towards the two performance essential parameters was due to the self-adjusting nature of it. Additionally, it indicates easier parameter setting for the user.

Limitations

The SAFD algorithm assumes Gaussian peak shapes in both mass and time domains, therefore, large deviations (e.g. irregular peak shapes) from this assumption may cause cases of false negatives. The SAFD algorithm assumes pure mass domain peaks hence the focus on the top 50% of the signal. A deviation from this assumption (i.e. mass resolution $\leq$ 10,000, based on the sensitivity analysis) may cause integration errors. As for the time domain, the features must have a chromatographic resolution of ≥ 0.75 for them to be detected by SAFD algorithm as two separate components.

The SAFD algorithm is computationally more expensive than other algorithms due to the fact that it uses all of the data points in the feature (i.e. the profile data) and it fits a 3D Gaussian into the data. For example, the tested dataset in the current study took SAFD around 7 hr to process versus 30 minutes with XCMS via MZmine. However, it should be noted that this is the first prototype of the algorithm and future optimizations may drastically decrease the run time.

Conclusions

SAFD is a robust, reliable, and accurate algorithm for non-targeted feature detection in the LC-HRMS profile data. This method takes advantage of all the measured points within

a feature without using arbitrary parameters. This algorithm has only two performance affecting parameters that are only used as a first guess or as an exception handling case. Consequently, it adjusts itself to fit the data in the best possible way. Therefore, SAFD, differently from the other methods, does not need any data binning, XIC generation, and/or RIO generation to perform feature detection. Therefore showing a great potential to be a widely used algorithm for non-targeted feature detection of profile LC-HRMS data.

Acknowledgement

We are thankful to the Research Council of Norway for the financial support of this project (RESOLVE, 243720). The authors are also grateful to Australian Research Council for financial support through project DP190102476.

Supporting Information Available

The Supporting Information including details regarding the chemicals, parameter settings of the algorithms, and figures related to each step taken within the SAFD algorithm is available free of charge on the ACS Publications website.

Author Information

Corresponding Author:

Saer Samanipour

E-mail: saer.samanipour@niva.no

Phone: +47 98 222 087

Address: Norwegian Institute for Water Research (NIVA)

Gaustadalléen 21, 0349 Oslo, Norway

References

(1) Schymanski, Emma L and Singer, Heinz P and Slobodnik, Jaroslav and Ipolyi, Ildiko M and Oswald, Peter and Krauss, Martin and Schulze, Tobias and Haglund, Peter and Letzel, Thomas and Grosse, Sylvia and others, *Anal. Bioanal. Chem.* **2015**, *407*, 6237–6255.

(2) Yi, L.; Dong, N.; Yun, Y.; Deng, B.; Ren, D.; Liu, S.; Liang, Y. *Anal. Chem. acta* **2016**, *914*, 17–34.

(3) Gorrochategui, E.; Jaumot, J.; Lacorte, S.; Tauler, R. *Trends Anal. Chem.* **2016**, *82*, 425–442.

(4) Samanipour, S.; Martin, J. W.; Lamoree, M. H.; Reid, M. J.; Thomas, K. V. *Environ. Sci. Technol.* **2019**,

(5) Samanipour, S.; Kaserzon, S.; Vijayasarathy, S.; Jiang, H.; Choi, P.; Reid, M. J.; Mueller, J. F.; Thomas, K. V. *Talanta* **2019**, *195*, 426–432.

(6) Parsons, B. A.; Pinkerton, D. K.; Wright, B. W.; Synovec, R. E. *J. Chromatogr. A* **2016**, *1440*, 179–190.

(7) Oberacher, H.; Arnhard, K. *TrAC Trends Anal. Chem.* **2016**, *84*, 94–105.

(8) Gago-Ferrero, P.; Schymanski, E. L.; Bletsou, A. A.; Aalizadeh, R.; Hollender, J.; Thomaidis, N. S. *Environ. Sci. Technol.* **2015**, *49*, 12333–12341.

(9) Chiaia-Hernandez, A. C.; Schymanski, E. L.; Kumar, P.; Singer, H. P.; Hollender, J. *Anal. Bioanal. Chem.* **2014**, *406*, 7323–7335.

(10) Samanipour, S.; Reid, M. J.; Bæk, K.; Thomas, K. V. *Environ. Sci. Technol.* **2018**, *52*, 4694–4701.

(11) Alygizakis, N. A.; Samanipour, S.; et al., *Environ. Sci. Technol.* **2018**, *52*, 5135–5144.

(12) Hollender, J.; Schymanski, E. L.; Singer, H. P.; Ferguson, P. L. *Environ. Sci. Technol.* **2017**,

(13) Samanipour, S.; Reid, M. J.; Thomas, K. V. *Anal. Chem.* **2017**, *89 (10)*, 5585–5591.

(14) Treviño, V.; Yañez-Garza, I.-L.; Rodriguez-López, C. E.; Urrea-López, R.; Garza-Rodriguez, M.-L.; Barrera-Saldaña, H.-A.; Tamez-Peña, J. G.; Winkler, R.; Díaz de-la Garza, R.-I. *J. Mass Spec.* **2015**, *50*, 165–174.

(15) Conley, C. J.; Smith, R.; Torgrip, R. J.; Taylor, R. M.; Tautenhahn, R.; Prince, J. T. *Bioinformatics* **2014**, *30*, 2636–2643.

(16) Tautenhahn, R.; Boettcher, C.; Neumann, S. *BMC Bioinformatics* **2008**, *9*, 504.

(17) Hites, R. A.; Jobst, K. J. Is Nontargeted Screening Reproducible? 2018.

(18) Smith, C. A.; Want, E. J.; O'Maille, G.; Abagyan, R.; Siuzdak, G. *Anal. Chem.* **2006**, *78*, 779–787.

(19) Pluskal, T.; Castillo, S.; Villar-Briones, A.; Orešič, M. *BMC Bioinformatics* **2010**, *11*, 395.

(20) Libiseller, G.; Dvorzak, M.; et al., *BMC Bioinformatics* **2015**, *16*, 118.

(21) Manier, S. K.; Keller, A.; Meyer, M. R. *Drug Test. Anal.* **2018**,

(22) Li, Z.; Lu, Y.; Guo, Y.; Cao, H.; Wang, Q.; Shui, W. *Anal. Chim. Acta* **2018**, *1029*, 50–57.

(23) Brodsky, L.; Moussaieff, A.; Shahaf, N.; Aharoni, A.; Rogachev, I. *Anal. Chem.* **2010**, *82*, 9177–9187.

(24) Coble, J. B.; Fraga, C. G. *J. Chromatogr. A* **2014**, *1358*, 155–164.

(25) Myers, O. D.; Sumner, S. J.; Li, S.; Barnes, S.; Du, X. *Anal. Chem.* **2017**, *89*, 8689–8695.

(26) Woldegebriel, M.; Derks, E. *Anal. Chem.* **2016**, *89*, 1212–1221.

(27) Woldegebriel, M.; Vivó-Truyols, G. *Anal. Chem.* **2015**, *87*, 7345–7355.

(28) Samanipour, S.; Baz-Lomba, N. A., Jose A. Alygizakis; Reid, M. J.; Thomaidis, N. S.; Thomas, K. V. *J. Chromatogra. A.* **2017**,

(29) Samanipour, S.; Langford, K.; Reid, M. J.; Thomas, K. V. *J. Chromatogra. A* **2016**, *1463*, 153–161.

(30) Samanipour, S.; Dimitriou-Christidis, P.; Nabi, D.; Arey, J. S. *Acs Omega* **2017**, *2*, 641–652.

(31) Keller, A.; Eng, J.; Zhang, N.; Li, X.-j.; Aebersold, R. *Mol. Syst Biol.* **2005**, *1*.

(32) Kessner, D.; Chambers, M.; Burke, R.; Agus, D.; Mallick, P. *Bioinformatics* **2008**, *24*, 2534–2536.

(33) Schoenberg, I. J. *Cardinal spline interpolation*; Siam, 1973; Vol. 12.

(34) Levenberg, K. *Q. App. Math.* **1944**, *2*, 164–168.

(35) Reynolds, D. *Encyclopedia of biometrics* **2015**, 827–832.

(36) Bezanson, J.; Karpinski, S.; Shah, V. B.; Edelman, A. *arXiv preprint arXiv:1209.5145* **2012**,

(37) Barrett, P.; Hunter, J.; Miller, J. T.; Hsu, J.-C.; Greenfield, P. matplotlib–A Portable Python Plotting Package. Astronomical data analysis software and systems XIV. 2005; p 91.

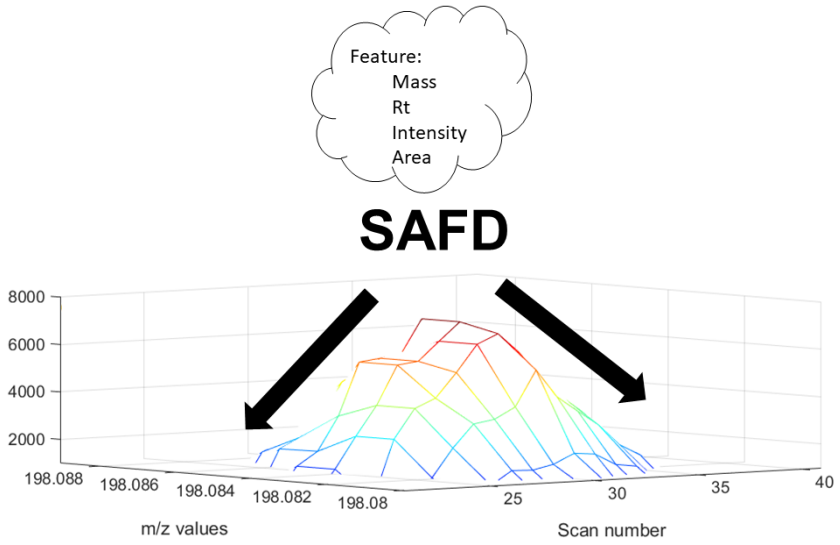
(38) Kiusalaas, J. *Numerical methods in engineering with Python 3*; Cambridge university press, 2013.

(39) Samanipour, S.; Baz-Lomba, J. A.; Reid, M. J.; Ciceri, E.; Rowland, S.; Nilsson, P.; Thomas, K. V. *Anal. Chem. acta* **2018**, *1025*, 92–98.

(40) Breslow, N. *Biometrika* **1970**, *57*, 579–594.

(41) Lawrence, I.; Lin, K. *Biometrics* **1989**, 255–268.

TOC Art



TOC Art for review only.

Supporting Information for: A Self Adjusting Algorithm for the Non-targeted Feature Detection of High Resolution Mass Spectrometry Coupled with Liquid Chromatography Profile Data

Saer Samanipour,[*,†,‡] Jake O´Brien,[‡] Malcolm J. Reid,[†] and Kevin V. Thomas[†,‡]

†*Norwegian Institute for Water Research (NIVA), Gaustadalléen 21, Oslo 0349, Norway*

‡*Queensland Alliance for Environmental Health Sciences (QAEHS), The University of Queensland, 20 Cornwall St, Woolloongabba, Qld 4102, Australia*

E-mail: saer.samanipour@niva.no

Phone: +47 22 18 51 00

Pages: 12

Figures: 13

Tables: 3

S1 Chemicals

Analytical grade formic acid was purchased from Sigma-Aldrich (Castle Hill, Australia). Analytical grade hydrochloric acid 32% was purchased from Univar (Ingleburn, Australia). Water was purified through a Milli-Q system. Liquid chromatography grade methanol was purchased from Merck (Darmstadt, Germany). High purity labelled internal standards were purchased from Novachem (Heidelberg West, Australia) with specific details listed in Table S1. Mobile phases were filtered using Sartorius Stedim 0.45 μm RC filters (Goettingen, Germany).

Table S1: The name, measured mass, and retention time of the internal standards (ISs).

nr	Name	m/z ($[M+H]^+$)[a] Da	Retention time[a] (min)
1	Atenolol-D7	274.214	3.18
2	Atrazine desethyl-D6	194.107	6.22
3	Atrazine desisopropyl-D5	179.085	4.37
4	Atrazine-D5	221.132	8.75
5	Caffeine	198.097	6.28
6	Carbamazepine-D10	247.165	9.11
7	Codeine-D3	303.179	5.18
8	Cotinine-D3	180.120	1.89
9	DEET-D7	199.182	9.35
10	Diuron-D6	239.061	8.76
11	Gabapentin-D10	182.195	3.49
12	Hexazinone-D6	260.084	7.54
13	Hydroxycotinine-D3	196.096	1.82
14	Imidacloprid-D4	260.084	7.54
15	Metolachlor-D6	290.179	10.17
16	Nicotine-D4	167.147	1.12
17	Paracetamol-D4	156.084	1.14
18	Simazine-D10	212.148	8.03
19	Venlafaxine-D6	284.251	7.78

[a] This is a measured value.

S2 Sample Treatment

The wastewater influent samples used for this study were collected as part of national sampling campaign in Australia where sample collection coincided with the 2016 Australian

Census.[1] Briefly, 24 hour composite samples were collected using existing onsite autosamplers operating in the optimized mode as outlined by Ort et al.[2] dependent on what was available at each site. Samples were aliquotted onsite into pre-cleaned (2 × methanol and 2 × MilliQ) HDPE bottles, had preservative added (samples used in this study were preserved with 2M HCl to adjust to ≈pH 2) and frozen prior to shipping frozen back to the lab. For this project, samples from 15 different WWTPs collected on Census day were chosen and covered a range of catchment sizes (from 3,500 people to more than 2.2 million people) and cover both metropolitan and regional places.

Prior to analysis, samples were defrosted, filtered with 0.2 μm RC filters (Phenomenex), 500 μL aliquotted into amber glass vials (Agilent 2 mL for LC) and 5 μL of a 1 μg/mL mix of internal standard (see SI for internal standards) added to each sample. A procedural blank and a QA/QC wastewater sample which continues to be analyzed with each batch of wastewater samples since 2016, were also prepared in the same way. An equilibrium sample consisting of just MilliQ without internal standards was also prepared. All samples, the blanks and the QA/QC were analyzed in triplicate but with the sequence in randomized order to prevent systematic error.

S3 LC-HRMS Conditions

Chemical analysis was performed on a Sciex 5600+ QToF (Sciex, Concord, Ontario, Canada) mass spectrometer with a DuoSpray Ion Source operating in positive electrospray ionization (ESI) mode coupled to a Shimadzu Nexera 2 HPLC system (Shimadzu Corp., Kyoto, Japan). Separation was achieved with a Kinetix Biphenyl column (2.6 μm, LC Column 50 mm × 2.1 mm, Phenomenex) at 45 °C using a mobile phase gradient of 5 to 100% methanol with 0.1% formic acid over a duration of 10 minutes with a mobile phase B curve of 2. The gradient was held at 100% B until 14.5 minutes before re-equilibrating to 5% until 17 minutes. A

pre-injection column (Altima C18 guard column) was used between the mobile phase and the injector to retard potential interferences from the mobile phase.

The mass spectrometer was operated in TOF MS mode with an accumulation time of 0.5 secs and a mass target range of 50 to 600 daltons. The ionization source was operated at 500 °C with an IonSpray Voltage of 5000 volts. Ion Source Gas 1 and Gas 2 were both set to 60 and Curtain Gas at 30. The Declustering Potential was set to 80 volts and the Collision Energy set to 10 volts. Calibration of the mass spectrometer was performed before analysis and after every fifth injection using the Sciex APCI Positive Calibration Solution: TOF MS delivered through a Calibrant Delivery System at a flow rate of 500 μL/min for 2 minutes.

S4 Algorithm Parameter Settings

Table S2: The parameter name, setting, function, and the comments related to the SAFD algorithm.

Parameter	Setting	Function	Comment
R^2	0.9	accept/reject Gaussian fit	Stopping parameter
Max Signal Increment	5[a]	Avoid overlapping peaks	Stopping parameter
Min Intensity	2000[b]	Defining baseline	Stopping parameter
Max Iteration	15,000	Max number of features	Stopping parameter
Min Peak Width	2[c]	Removing noise	Filtering parameter
Max Peak Width	300[c]	Removing noise	Filtering parameter
Resolution	20,000	The first guess	Performance parameter
Min Mass Peak Width	0.02[d]	Exception handling	Performance parameter

[a] This parameter is in % signal increment; [b] The unit for this parameter is counts (or absolute signal intensity); [c] The unit of this parameter is seconds; [d] The unit for this parameter is Da.

Table S3: The parameter name, setting, function, and the comments related to the XCMS via MZmine algorithm.

Parameter	Setting	Function	Comment
Noise level	2000	Defining baseline	Mass detection
Scale level	20[a]	Defining significant peaks	Mass detection
Wavelet window	5[a],[b]	Peak width mass domain	Mass detection
Min time span	0.02[c]	Min peak time domain	Chromatogarm builder
Min height	2000[d]	Removing noise	Chromatogarm builder
m/z tolerance	0.01[e]-20[f]	Grouping masses in XIC	Chromatogarm builder
Wavelet scales	0.1-2[g]	Peak detection in XIC	Deconvolution
Peak duration	0.02-3[a],[c]	Peak width time	Deconvolution

[a] This parameter was optimized using show preview function; [b] The unit for this parameter is %; [c] The unit of this parameter is minutes; [d] The unit for this parameter is counts (i.e. absolute signal intensity); [e] This parameter is expressed in Da; [f] The unit for this parameter is ppm; [g] The unit for this parameter is minutes.

S5 SAFD Algorithm

Figures S1, S2, S3, S4, S5, S6, S7, and S8 are showing all the steps taken by SAFD algorithm during each iteration (i.e. the detection of one feature).

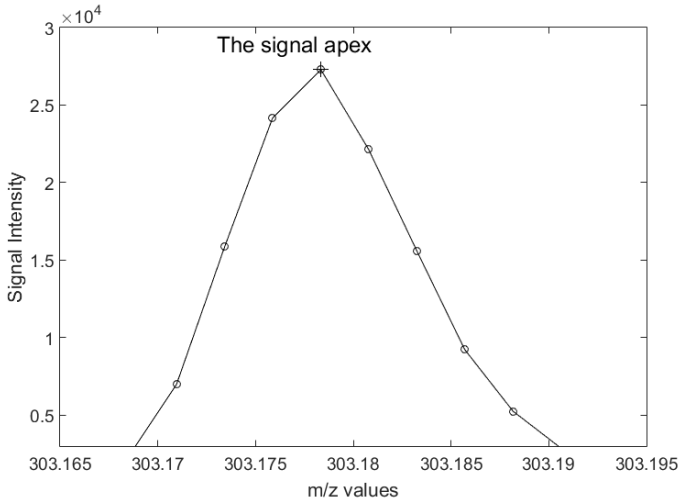


Figure S1: Depicting the maximum detection of a peak in the mass domain, step 1. The presented plot is based on a feature of an IS in the wastewater influent sample.

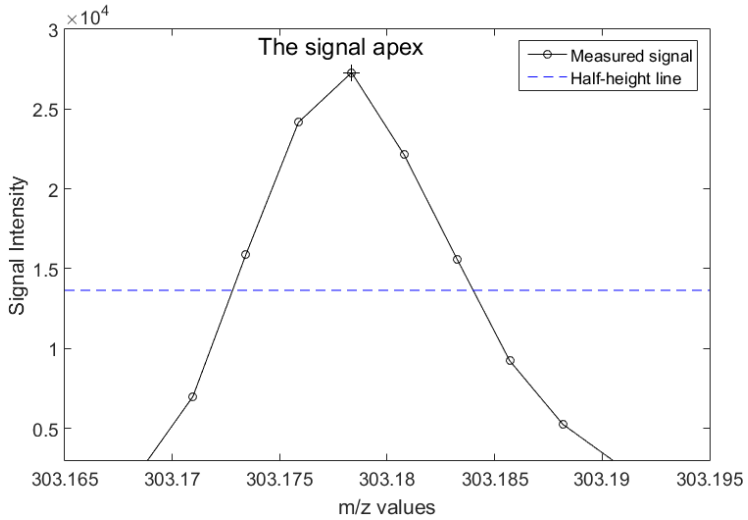


Figure S2: Depicting the detection of the half-height of a peak in the mass domain, step 2. The presented plot is based on a feature of an IS in the wastewater influent sample.

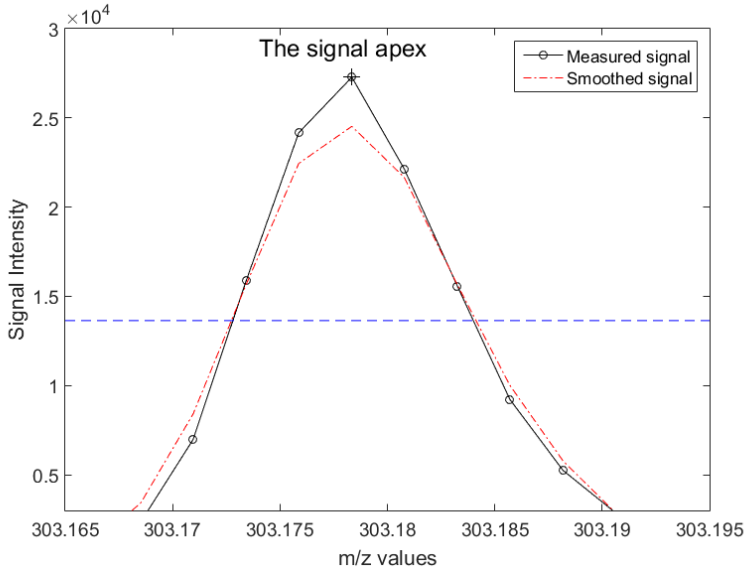


Figure S3: Depicting the process of smoothing a peak in the mass domain using the moving average method with a window of 3 points, step 3. The presented plot is based on a feature of an IS in the wastewater influent sample.

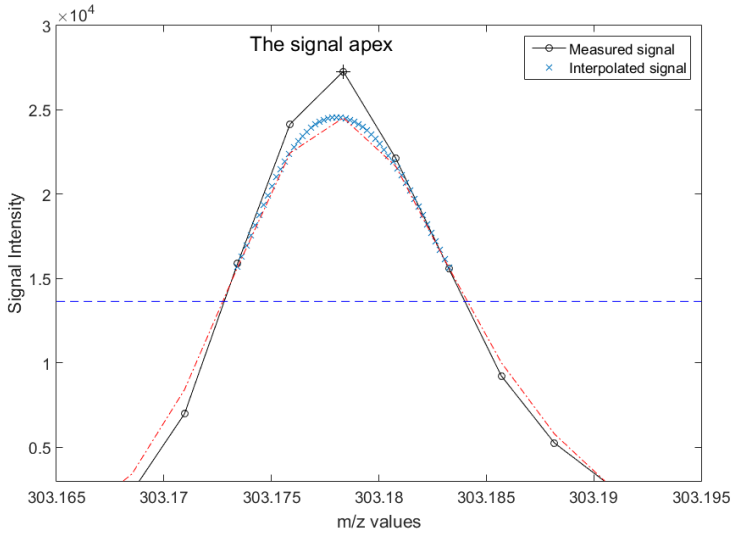


Figure S4: Depicting the interpolation of the smoothed signal using the Spline function,[3] step 4. The presented plot is based on a feature of an IS in the wastewater influent sample.

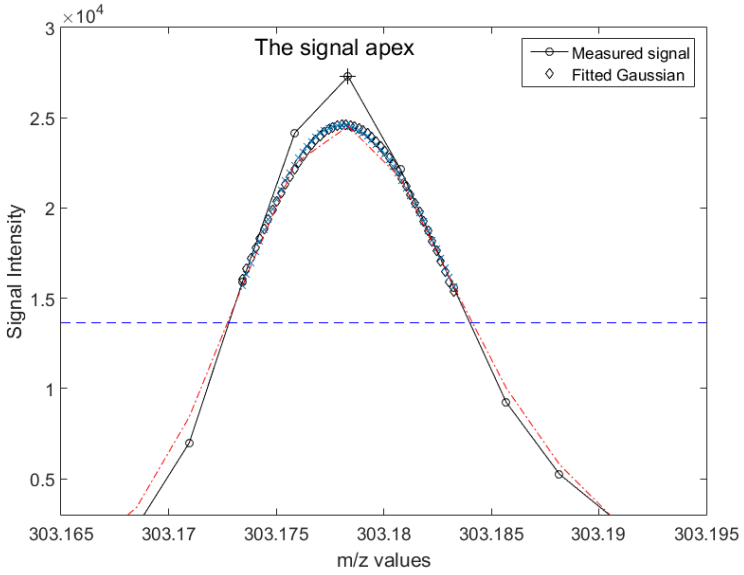


Figure S5: Showing the interpolated signal fitted by a Gaussian function via least square method,[4] step 5. The presented plot is based on a feature of an IS in the wastewater influent sample.

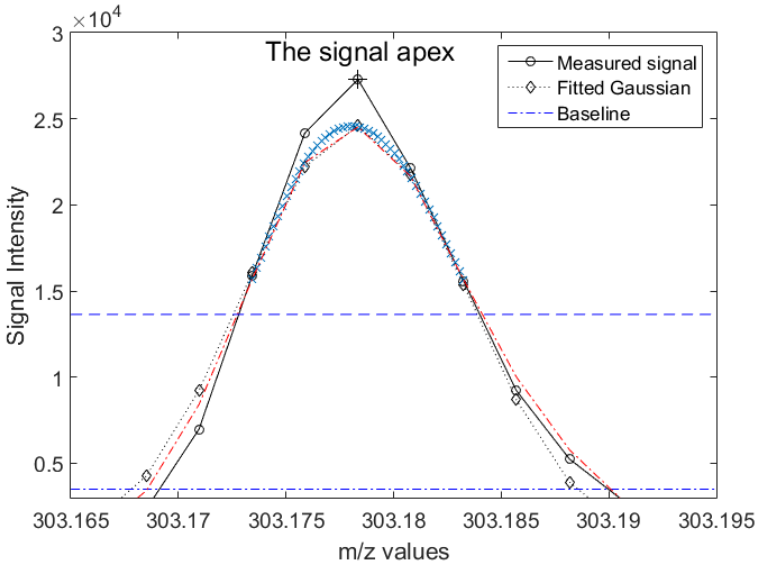


Figure S6: Depicts the process of tracing the baseline (i.e. minimum signal intensity) in the real signal through the fitted Gaussian function, step 6. The presented plot is based on a feature of an IS in the wastewater influent sample.

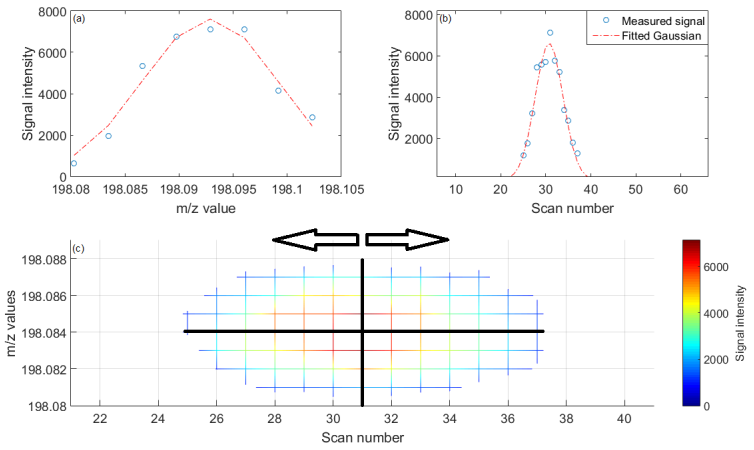


Figure S7: Shows (a) the fitted Gaussian on the base peak in the mass domain, (b) the fitted Gaussian in the time domain, and (c) a 3D overview of algorithm moving one scan at the time from the base peak in the mass domain (i.e. the black vertical line) to the neighboring scans in both directions. In panel (c) each vertical line and the horizontal black line represent a fitted Gaussian, step 7. The presented plot is based on a feature of an IS in the wastewater influent sample.

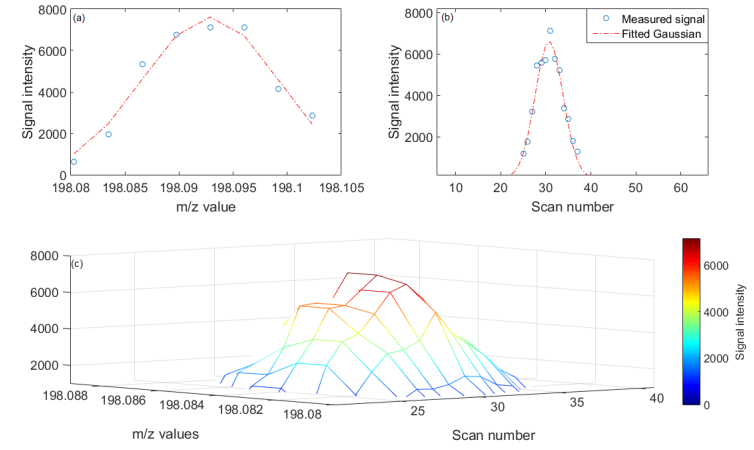


Figure S8: Depicts (a) the fitted Gaussian on the base peak in the mass domain, (b) the fitted Gaussian in the time domain, and (c) a 3D plot of the detected feature that will be set to baseline (i.e. minimum signal intensity), step . The presented plot is based on a feature of an IS in the wastewater influent sample.

S6 Feature Detection via SAFD Algorithm

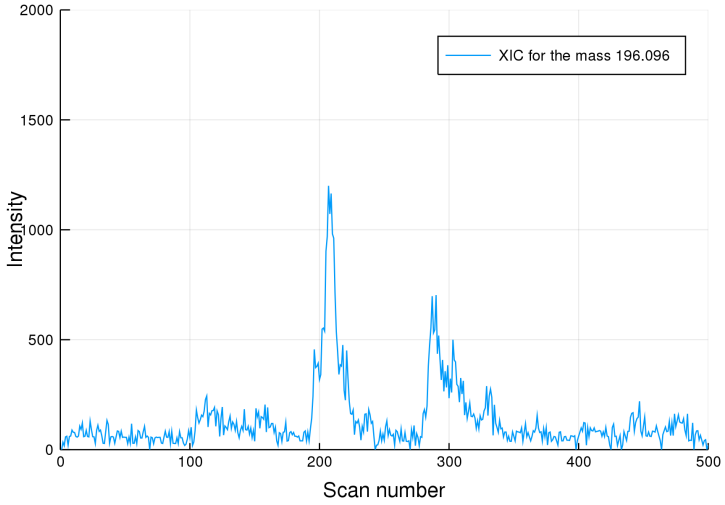


Figure S9: Shows the extracted ion chromatogram of hydroxycotinine in one of the samples, which was one of the FNs due to lower intensity of the base peak than the minimum signal intensity of 2000 counts.

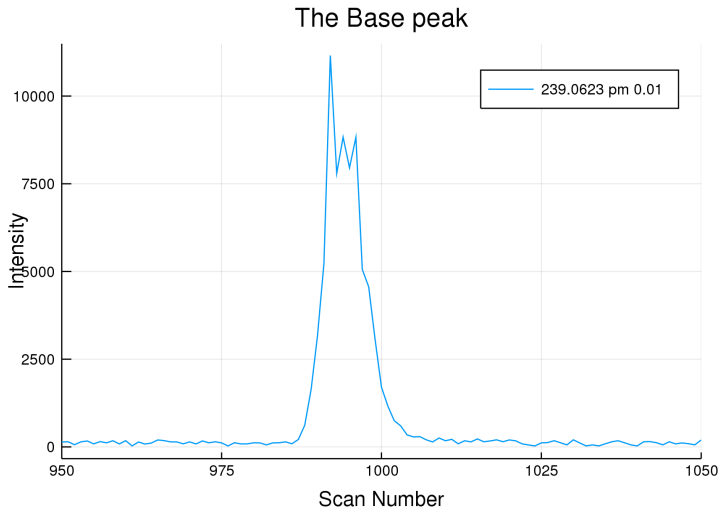


Figure S10: Depicts the extracted ion chromatogram of diuron, which was one of the four cases of FNs due to the high level of noise in the time domain.

S7 Feature Integration

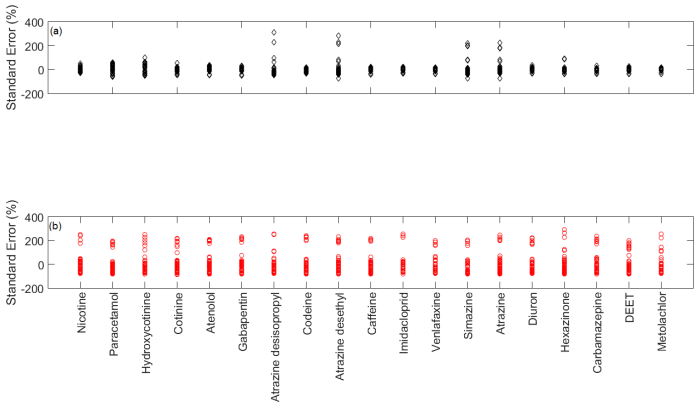


Figure S11: Shows the calculated average absolute standard error of the feature area for each IS over 51 spiked samples via (a) SAFD algorithm and (b) XCMS implemented through MZmine.

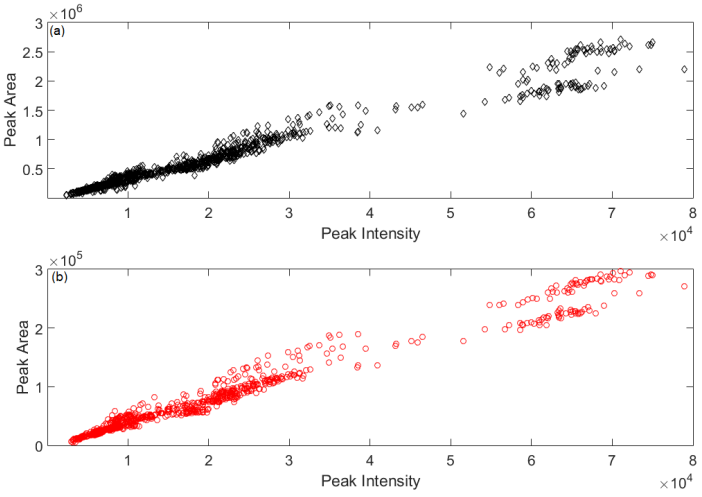


Figure S12: Depicts the feature height (i.e. intensity) vs feature area via (a) SAFD algorithm and (b) XCMS implemented through MZmine.

S8 Sensitivity Analyses for SAFD Algorithm

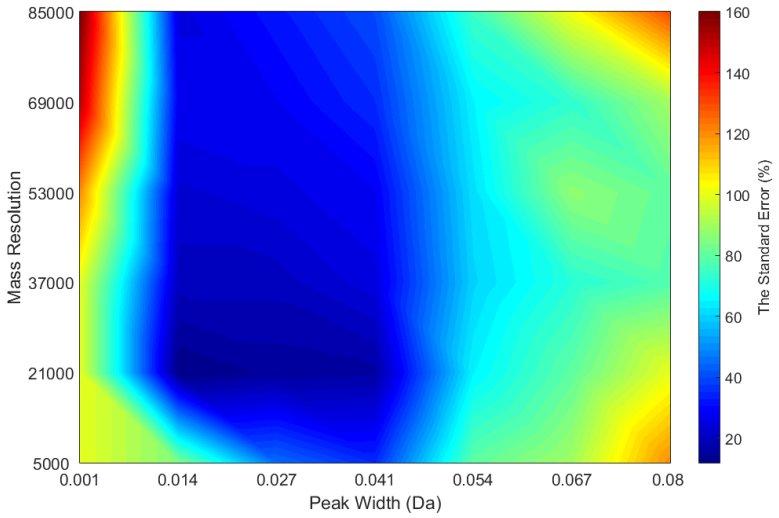


Figure S13: Shows the averaged absolute standard error of feature area as a function of the peak width and mass resolution parameters.

References

(1) O'Brien, J.; Grant, S.; Banks, A.; Bruno, R.; Carter, S.; Choi, P.; et al, *Environ. Int.* **2018**,

(2) Ort, C.; Lawrence, M.; Rieckermann, J.; Joss, A. *Environ. Sci. Technol.* **2010**, *44 (16)*, 6024–6035.

(3) Schoenberg, I. J. *Cardinal spline interpolation*; Siam, 1973; Vol. 12.

(4) Levenberg, K. *Q. App. Math.* **1944**, *2*, 164–168.